FREEtree: A Model Tree Based Approach for Longitudinal and Clustered Data with Correlated Features

Yuancheng Xu[1], Dan Kojis[2], Min Tan[3], Christina Ramirez[4,*]

[1]Department of Mathematics, Southern University of Science and Technology; [2]Department of Statistics, University of Wisconsin–Madison; [3]Department of Mathematics, Sichuan University; [4]UCLA Fielding School of Public Health, Department of Biostatistics, University of California, Los Angeles

Introduction

Longitudinal data means that repeated observations over time are available for each object (such as a patient) and clustered data, a more general situation, means that observations are nested in a hierarchical structure within objects.

Challenges:

- Correlated observations within an object
- Large number of features: Select features to improve interpretability
- Highly correlated features: When selecting features, most classical machine learning methods are biased towards correlated features and often ignore independent ones.

Fuzzy Random Effect Estimation tree (FREEtree) deals with correlated observations (longitudinal or clustered structure) using piecewise *Mixed-Effect Models* and provide relatively unbiased feature selection in the presence of highly correlated features by a *screening and selection* step. Finally the selected features are used to train a model tree for prediction.

FREEtree Algorithm

Building Block: Linear Mixed-Effect Model Tree (LMM tree[1])

LMM tree is a piecewise linear mixed-effect model function as follows. Only random intercepts are included here for illustration, though random slopes are allowed. Splitting and regression features can be specified by users. LMM model tree is trained within the framework of the expectation-maximization (EM) algorithm.

$$\hat{y}_{it} = \begin{cases} \beta_1^T \tilde{x}_{it} + C_1 \\ \beta_2^T \tilde{x}_{it} + C_2 \\ \beta_3^T \tilde{x}_{it} + C_3 \end{cases} + b_i$$

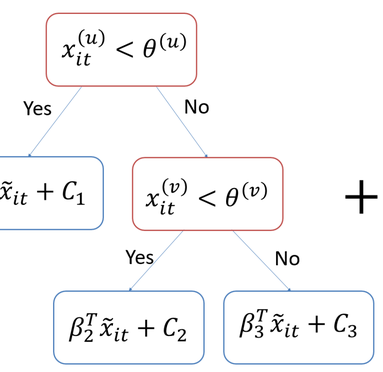

Fig 1: LMM tree. i is the index of object and t is the index of measurement within an object. $x_{it}^{(u)}$ is splitting feature u and $\tilde{x}_{it}$ is subset of features used for regression. b_i is the random intercept of object i.

Feature Selection & Prediction:

Motivated by Fuzzy Forest[2], we first cluster features using Weighted Correlation Network Analysis (WGCNA) so that correlation is large within a non-grey module and small between modules. The grey module P0 contains relatively independent features. Then screen features within non-grey modules regressing on the dominant *principal components* (PCs) if regressors are not specified. Selection from screened non-grey features follows and the selected ones are used as regressors to screen grey features. The final features are used both as regressors and splitters in LMM tree.

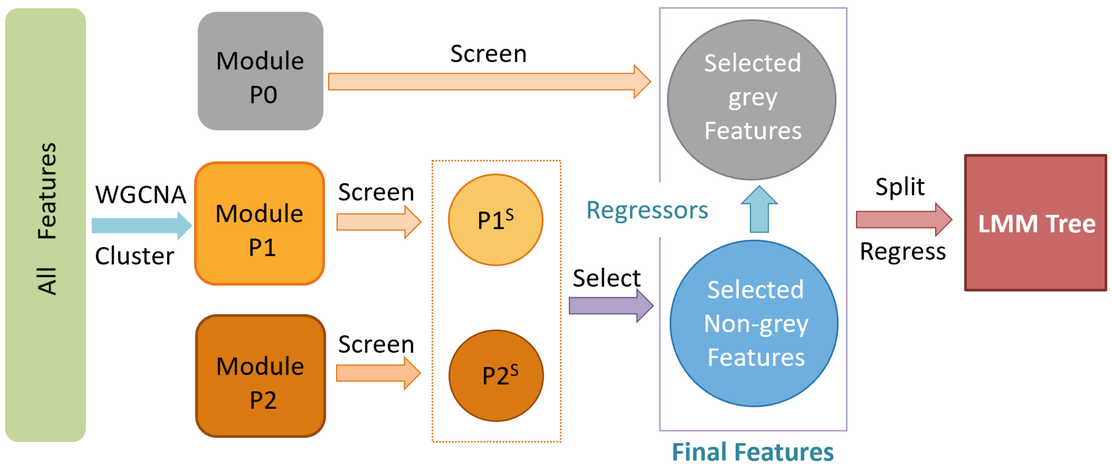

Fig 2: Flow chart of FREEtree

Simulations and Results

Features 1-100,101-200,201-300 and 301-400 are grouped with the last module being independent and the first three correlated within each module. Features from different modules are independent. The first dataset is generated using the following process where f only relates to feature 1,2,3,301,302,303 and includes *time-treatment interaction*. The second dataset does not include this interaction and PCs are used as regressors when screening non-grey features.

$$y_{it} = f(X_{it}) + (t-3)^2 \mathbb{1}_{treatment1} - (t-3)^2 \mathbb{1}_{treatment2} + b_i + \epsilon_{it}$$

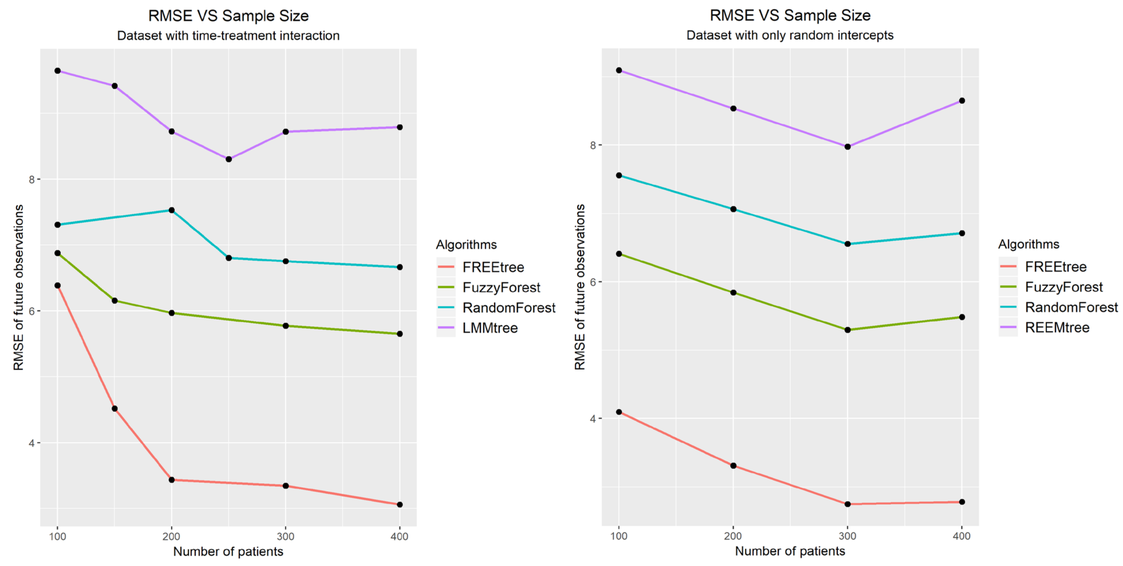

Fig 3,3': Predictive Performance

Sample size	Important features	Other features
100	1,2,3,301,302	NA
150	1,2,3,301,302,303	53,94
200	1,2,3,301,302,303	NA
300	1,2,3,301,302,303	NA
400	1,2,3,301,302,303	368

Table 1: Feature selection performance on the time-treatment dataset. When sample size is larger than 150, all important features are correctly identified.

Sample Size	treatment1		treatment2	
	time	time2	time	time2
100	5.23	1.36	−8.88	−0.89
200	5.46	0.88	−5.60	−0.91
300	5.43	0.99	−6.06	−0.91
400	6.16	1.07	−6.40	−1.01

Table 2: Estimation of underlying structure on time-treatment dataset. The coefficient of t and t^2 with different treatments are correctly estimated.

Conclusion and Future Work

FREEtree provides a framework to do feature selection and prediction on longitudinal and clustered data. By using LMM tree, FREEtree has more fitting and predictive power than regular tree such as CART and RE-EM tree[3]. By adopting a screening-selecting procedure and the use of dominant principal components as regressors when screening , FREEtree reduces the bias in feature selection and dimension of feature space, resulting in a interpretable model tree.

Future work includes dealing with covariates that also have time structure. In this case, WGCNA should be adapted to cluster time series. One potential solution is to replace correlation with time series distance measure such as dynamic time warping (DTW) although computational cost will be the main concern.

References

1.Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. Behavior research methods, 50(5), 2016-2034. 2. Conn, D., Ngun, T., Li, G., & Ramirez, C. (2015). Fuzzy forests: extending random forests for correlated, high-dimensional data. 3. Sela, Rebecca J., and Jeffrey S. Simonoff. "RE-EM trees: a data mining approach for longitudinal and clustered data." Machine learning 86.2 (2012): 169-207.